DOCUMENT RESUME

ED 386 473 TM 023 976

AUTHOR Curley, W. Edward; Schmitt, Alicia P.
TITLE Revising SAT-Verbal Items To Eliminate Differential Item Functioning. College Board Report No. 93-2.
INSTITUTION College Board, New York, NY.; College Entrance Examination Board, New York, N.Y.
REPORT NO ETS-RR-93-61
PUB DATE 93
NOTE 23p.; Version of a paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
AVAILABLE FROM College Board Publications, Box 886, New York, NY 10101-0886 ($12).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Entrance Examinations; High Schools; *High School Students; *Item Bias; *Minority Groups; *Predictive Validity; *Test Construction; Test Items; Verbal Tests
IDENTIFIERS Mantel Haenszel Procedure; *Scholastic Aptitude Test

ABSTRACT

Based on initial Scholastic Aptitude Test (SAT) Verbal pretest data and hypotheses advanced in the research literature, 7 sentence completion and 16 analogy items with extreme levels of differential item functioning (DIF) were selected and then systematically revised and re-administered in an attempt to reduce or eliminate DIF. The apparent success of the effort makes similar attempts worth continuing. The particular terminology used in stems and keys, rather than the underlying skill being measured, seemed to be a recurring source of DIF in the SAT-Verbal items. Larger sample sizes, especially for minority focal groups, would help to stabilize the DIF categories used by Educational Testing Service (ETS) test developers. In addition, because the ETS delta metric is unbounded at the extremes, the use of both the Standardization (p-metric) and Mantel-Haenszel (delta-metric) methodologies is recommended for classifying the level of DIF for very easy and very difficult items. Further research is suggested to study the possible relationship between DIF and predictive validity. Nine tables and four figures present analysis results. An appendix summarizes DIF hypotheses. (Contains 26 references.) (Author/SLD)

College Board Report No. 93-2

Revising SAT®-Verbal Items to Eliminate Differential Item Functioning

W. EDWARD CURLEY and ALICIA P. SCHMITT

The College Board
Educational Excellence for *All* Students

College Board Report No. 93-2
ETS RR No. 93-61

Revising SAT®-Verbal Items to Eliminate Differential Item Functioning

W. EDWARD CURLEY and ALICIA P. SCHMITT

College Entrance Examination Board, New York, 1993

Acknowledgments

A prior version of this paper was presented at the annual meeting of the American Educational Research Association/ National Council on Measurement in Education, San Francisco, April 1992.

Funding was provided by the College Board/ETS Joint Staff Research and Development Committee.

Special thanks are extended to Kathy Hassler for her fine efforts in typing the manuscript and formatting the tables.

W. Edward Curley is a senior examiner in Test Development at ETS.

Alicia P. Schmitt is a senior measurement statistician at ETS.

The College Board is a national nonprofit association that champions educational excellence for all students through the ongoing collaboration of more than 2,900 member schools, colleges, universities, education systems, and organizations. The Board promotes—by means of responsive forums, research, programs, and policy development—universal access to high standards of learning, equity of opportunity, and sufficient financial support so that every student is prepared for success in college and work.

Additional copies of this report may be obtained from College Board Publications, Box 886, New York, New York 10101-0886. The price is $12.

Printed in the United States of America.

Contents

Abstract 1

Introduction 1

Method 2
Data Source 2
Instrument and Design 2
Procedure 3

Results and Discussion 4
Two Initial Observations 7
Success Rate in Reducing Differential Item Functioning 9
Factors Related to Differential Item Functioning 10

Conclusions 15

References 16

Appendix: Summary of Hypotheses about DIF Relevant to the SAT-Verbal Items Selected for this Study 18

Figures
1. Scatterplot of difficulty estimates between pairs of items: Forms A and B. 5
2. Scatterplot of discrimination estimates between pairs of items: Forms A and B. 5
3. Scatterplot of difficulty estimates between pairs of items: Forms C and D. 5
4. Scatterplot of discrimination estimates between pairs of items: Forms C and D. 5

Tables
1. Sample Sizes and Difficulty Estimates (P%) for Study Groups across Forms A, B, C, and D 2
2. Effects of Science Terminology 6
3. Effects of Industrial Arts Terminology 7
4. Effects of Military Terminology 8–9
5. Effects of Contexts Portraying Aggression or Conflict 11
6. Effects of Special Interest Terminology 12
7. Effects of Cognates 12
8. Effects of Homographs 13
9. Mantel-Haenszel Values (MH D-DIF) and ETS DIF Categories for Selected SAT-Verbal Items Reprinted in this Study Identically to the Initial Pretest 14

Abstract

Based on initial SAT-Verbal pretest data and/or hypotheses advanced in the research literature, the authors selected 7 sentence completions and 16 analogies with extreme levels of differential item functioning (DIF) and then systematically revised and readministered the items in an attempt to reduce or eliminate DIF. Several diverse conclusions can be drawn from the data. First, because of the apparent success in reducing extreme levels of DIF in SAT-Verbal items, the authors recommend that such efforts be continued. Second, the particular terminology used in stems and keys (rather than the underlying reasoning skill being measured) seems to be a recurring source of DIF in SAT-Verbal items. Third, larger sample sizes, particularly for minority focal groups, would help to stabilize the DIF categories used by Educational Testing Service (ETS) test developers. Fourth, because the ETS delta metric is unbounded at the extremes, the use of both the Standardization (p-metric) and Mantel-Haenszel (delta-metric) methodologies is recommended for classifying the level of DIF for very easy and very difficult items. Finally, the paper concludes with a suggestion for further research concerning the possible relationship between DIF and predictive validity.

Introduction

Differential item functioning (DIF) statistics can be used to identify test questions on which the various focal (minority or female) and reference (white or male) populations perform differently. Since the mid-1980s, a series of DIF studies on the operational verbal sections of the SAT has been conducted to identify and assess the nature of the items on which DIF can be observed (Schmitt 1985; Bleistein and Wright 1986; Wendler and Carlton 1987; Rogers and Kulick 1987; Schmitt and Bleistein 1987; Schmitt 1988; Lawrence, Curley, and McHale 1988; Lawrence and Curley 1989; Schmitt and Dorans 1990).

In addition, randomized studies of specially constructed items have been undertaken in an attempt to isolate and evaluate factors—both within and across item types and testing programs—that may consistently result in elevated levels of DIF for one or more focal groups (Scheuneman 1987; Dorans, Schmitt, and Curley 1988; Scheuneman and Briel 1988; Schmitt, Curley, Bleistein, and Dorans 1988; Bleistein, Schmitt, and Curley 1990). This latter group of studies has generally used items written specifically for the study or pretested items on which no DIF data were yet available; these items have been administered in nonoperational sections that did not count as part of the examinees' scores.

Although findings from the randomized studies have clarified some factors previously hypothesized to be related to DIF, they have also shown that elevated levels of DIF cannot be completely eliminated at the item-writing stage because the factors are confounded or as yet unidentified. The mere flagging of an item for DIF does not indicate the reason(s) for the differential functioning. Thus it is likely that some SAT items with elevated levels of DIF will continue to be found at the pretest stage even if test developers were provided with item-writing guidelines and/or if changes were made in test specifications to reduce DIF. Of the approximately 2,250 SAT-Verbal (SAT-V) questions pretested during 1990, 190 items (about 8.5 percent) exhibited moderate to large amounts of DIF. This total includes items differentially advantaging, as well as disadvantaging, focal groups, and includes questions from all four of the verbal item types (antonyms, analogies, sentence completions, and reading comprehension).

Elevated levels of DIF in and of themselves do not prove that test questions are biased. Once an item is flagged for high DIF, judgment should be used to decide whether the difference in difficulty shown by the DIF index is unfairly related to group membership. The determination of fairness should be based on whether or not the difference in difficulty is judged to be relevant to the construct being measured by the test (Zieky 1991). For the purposes of this study, the authors selected pretested items with elevated levels of DIF that had not yet been evaluated with respect to whether or not the DIF was construct-relevant. The majority of the items studied contained specialized terminology likely to be found in the material read and viewed, or in the language used, or in the experiences engaged in, by one gender or minority group more often than by another because of their particular interests or opportunities.

The chief purposes of this study were twofold: (1) to revise individual verbal items on which elevated levels of DIF had been observed at the pretest stage to try to reduce or eliminate the DIF and thus make the items appropriate for use in operational forms of the SAT; and (2) to continue to evaluate and perhaps to supplement hypothesized DIF factors for the SAT by observing the effects of revisions on individual items previously exhibiting DIF.

Method

Data Source

The data for this study were collected (1) initially in pretest sections at various regular Saturday administrations of the SAT and then, after the selected items were systematically revised and reassembled into four 30-minute nonoperational sections, (2) at a regular 1991 SAT administration. At both stages of data collection the data consisted of unscored item responses to nonoperational questions from random samples of self-reported females, males, Asian Americans, blacks, Hispanics, and whites for whom English was (or English and another language were) their first language(s). Sample sizes for the analyses of the four newly assembled forms (labeled A to D) ranged from 4,331 white examinees for Form A to 183 Hispanic examinees for both Forms C and D (see Table 1). Data from the earlier, initial pretesting of the selected items were based on groups of examinees that were roughly proportional in sample size to the groups reported in Table 1.

TABLE 1

Sample Sizes and Difficulty Estimates (P%) for Study Groups across Forms A, B, C, and D

	Form A	*Form B*	*Form C*	*Form D*
		WHITE		
N	4,331	4,112	4,061	3,856
Mean P%	56	58	60	56
S.D. P%	26	23	22	25
		HISPANIC		
N	199	235	183	183
Mean P%	45	50	49	49
S.D. P%	24	21	21	24
		BLACK		
N	621	573	575	565
Mean P%	39	41	43	41
S.D. P%	23	21	21	21
		ASIAN AMERICAN		
N	259	270	237	226
Mean P%	60	59	65	60
S.D. P%	23	22	20	24
		MALES		
N	2,511	2,582	2,488	2,311
Mean P%	55	56	59	55
S.D. P%	25	23	22	24
		FEMALES		
N	3,028	2,746	2,689	2,625
Mean P%	53	56	57	54
S.D. P%	25	23	22	25

Instrument and Design

Based on initial pretest DIF data and/or hypotheses proposed in previous research (see the Appendix), a total of 7 sentence completion and 16 analogy items were selected as the focus of investigation. These items were then modified in ways intended to eliminate the factors hypothesized to be related to the elevated levels of DIF observed at the initial pretesting. Whenever possible, given the available "pool" of SAT items showing moderate to large amounts of DIF, sets of two or more items with the same (or similar) hypothesized DIF factors were included in this study for purposes of internal replication. The different versions of each of the 23 items studied are displayed in Tables 2 to 8 in terms of the various hypothesized DIF factors.

Seven DIF factors were examined. Not all (or even most) items in the following seven general categories consistently show elevated levels of DIF, but certain patterns have been detected. Science, industrial arts, and military terminology, as well as contexts portraying aggression or conflict, may negatively affect the performance of females, based on what has been found in the evaluation of some SAT-V items and/or in the research literature. Terminology of special interest or familiarity to particular groups may positively affect the performance of those groups. Cognates with Spanish, especially when they appear in the stem or key of SAT-V questions, may positively affect the performance of Hispanic examinees. Homographs, especially when they appear in the stem or key of SAT-V questions, may negatively affect the performance of Hispanic, black, and Asian American examinees. These are the seven DIF factors examined in the present investigation.

The original versions of the items studied (worded identically to the initial pretests) and as many as three different revised versions of each were assembled into four sections of the SAT for re-pretesting. Each section consisted of 40 verbal questions presented in the same order as that of the 40-item operational section of the SAT-V: items 1 to 10 were identical antonym questions across the four forms and were not part of this investigation; items 11 to 15 were sentence completions; items 16 to 25 were analogies; and items 26 to 40 were reading comprehension questions and not part of this investigation. Thus, Forms A, B, C, and D were indistinguishable from the operational sections of the SAT-V (as were the earlier verbal pretests from which the initial DIF data were derived).

Original and revised versions of items were kept in the same position across either two or four of Forms A,

B, C, and D. For example, item 11 in Forms A, B, C, and D presented four different versions of the same sentence completion; item 12 in Forms A and B presented two different versions of a second sentence completion; item 12 in Forms C and D presented two different versions of a third sentence completion; and so on. Each of the four forms was constructed in such a way that it would not violate any of the usual SAT-V pretest assembly guidelines. To the extent possible, the difficulty of the alternate versions of the items was kept parallel; however, insofar as word substitutions were based on the subjective judgments of the authors, alternate versions of some items were found to differ in difficulty.

Variables such as order of answer choices (A to E), key position, and content classification (unless it was associated with the hypothesis being evaluated) were held constant for the alternate versions of each item studied. The factors hypothesized as causes of the elevated DIF as well as the various groups differentially affected by the items studied were also carefully balanced across the four forms so that no one section of re-pretested items would include a preponderance of questions likely to affect any particular group either negatively or positively.

In addition to reviews by the authors, each item studied and its alternate version(s) were also reviewed by two test development colleagues familiar with the SAT-V and with relevant DIF research. After the pretested items were assembled into the four sections for this study, each of the variants passed through routine test specialist, editing, sensitivity, and planograph reviews. This review process assured that the four nonoperational forms from which data for this investigation were derived were comparable to regular operational and pretest sections of the SAT-V.

Procedure

Items that are more difficult for one group than for another with the same level of ability or skill are defined as differentially more difficult or as functioning differentially between the two groups. Usually the white or male group is referred to as the reference or base group and the minority or female group as the focal or study group. Since DIF indices take into account overall differences in ability on the construct being measured by matching the groups before comparing their performance, DIF indices identify items that might have construct-irrelevant characteristics.

Two statistical procedures currently used at ETS to assess DIF are the Mantel-Haenszel (MH) method (Holland and Thayer 1988) and the Standardization (DSTD) method (Dorans and Kulick 1983, 1986). Both of these methods identify DIF after partitioning the reference and focal groups into subgroups with the same score on a relevant matching variable. The matching variable is usually the total score on a test closely related to the construct that the item is intended to measure. While there are some minor differences between the MH and DSTD methods (Dorans and Holland 1992), the DIF estimates computed by these methods are highly correlated (in the upper .90s) because they tend to yield the same rank order of items with respect to DIF (Wright 1987; Holland and Thayer 1988; Dorans 1989). Both the DSTD and MH indices take into account speededness in the calculations of DIF by including only those examinees who reached an item in the calculation of the DIF value for that item (Schmitt and Bleistein 1987).

Standardization Procedure

In the traditional Standardization analysis, an item is said to exhibit differential item functioning when the probability of correctly answering the item is lower or higher for examinees from one group than for equally able examinees from another group. The focus of DIF analyses is on differences in performance between groups that are matched with respect to the ability, knowledge, or skill of interest.

The basic elements of a Standardization analysis of the keyed response are proportions correct at each level of a matching variable, such as total score, in a base or reference group and a focal or study group. Standardization provides the DSTD index for quantifying DIF in the p metric. This index can range from −1 to +1, or from −100 percent to 100 percent. Negative values of DSTD indicate that the item disadvantages the focal group, while positive values indicate that the item favors the focal group. STD P- DIF values between −.05 (−5 percent) and +.05 (+5 percent) are considered negligible. STD P-DIF values outside the −.10 and +.10 (or the −10 percent, +10 percent) range are considered sizable. For operational purposes, a $|DSTD| \geq .10$ is a recommended cutoff; for exploratory research purposes, a less reliable cutoff of $|DSTD| \geq .05$ is often used. In addition to calculating DSTD values for the key, differences in the standardized proportion of responses for each distractor are also computed and studied to understand better the effects of the hypothesized DIF factors (see Dorans, Schmitt, and Bleistein, 1992, for a description of distractor analyses).

Mantel-Haenszel Method

The Mantel-Haenszel procedure (Mantel and Haenszel 1959), adapted by Holland and Thayer (1988) for DIF analysis, computes ratios of the conditional odds of successful reference group performance over the conditional odds of successful focal group performance at each score

level, and then averages these ratios across score levels. In the calculation of the average ratio, statistically optimal weights are used for each ratio. The Mantel-Haenszel method provides an estimate of the constant odds-ratio.

The MH statistic is transformed to the "delta" metric used to indicate item difficulty in the ETS test development process. To obtain a delta, the proportion correct (p) is converted to a z-score via a p-to-z transformation using the inverse of the normal cumulative function, followed by a linear transformation to a metric with a mean of 13 and a standard deviation of 4. Large values in a delta metric correspond to difficult items, while easy items have small delta values. This MH estimate of DIF effect size in the delta metric ranges from negative infinity to infinity, with a value of 0 indicating no DIF. MH D-DIF values between −1.00 and +1.00 are considered negligible. MH D-DIF values outside the −1.50, +1.50 range are considered sizable. For operational purposes, |MH D-DIF|≥1.50 is a recommended cutoff; a less reliable cutoff of |MH D-DIF|≥1.00 is often used for exploratory research purposes. As with DSTD, positive values of MH D-DIF favor the focal group, while negative values disadvantage the focal group. For a complete description and comparison of the DSTD and MH D-DIF statistics, refer to Dorans and Holland (1992).

In the present investigation, categorization of DIF items was made on the basis of the standard ETS DIF operational item screening classifications (Petersen 1988). These classifications are as follows:

(1) "A" items have a MH D-DIF *not* significantly different from 0 (at the .05 level) or an absolute value less than 1.00; (2) "B" items have a MH D-DIF significantly different from 0 (at the .05 level) *and* either an absolute value of at least 1.00 but less than 1.50 *or* an absolute value of at least 1.00 but not significantly greater than 1.00 (at the .05 level); (3) "C" items have an absolute value of MH D-DIF of at least 1.50 *and* significantly greater than 1.00 (at the .05 level).

Matching Criteria

The analysis of differential item functioning involves a two-step process to refine the matching criteria. During the first step, the total-test raw score on the SAT-V operational 85-item test is used as the matching criterion to determine DIF for each of the 85 items. On the basis of the initial analysis, any item with extreme DIF values for the corresponding focal group comparison is removed as part of the total score used to match the reference and focal groups. Thus, a "refined" matching criterion is determined for each focal group comparison for use in the subsequent pretest DIF analyses. In this study, two items were identified as having extreme DIF (one for both black and Asian American examinees and one for only Asian American examinees) and were, therefore, deleted from the total score matching criterion for the respective focal group analyses. Thus two refined matching criteria were created: (1) for the white and Asian American examinees: SAT-V = 83 (85 items minus 2 items) and (2) for the white and black examinees: SAT-V = 84 (85 items minus 1 item). For the other focal groups (i.e., Hispanic and female examinees), the matching criterion was the total score on the 85-item SAT-V operational test.

Results and Discussion

Difficulty estimates for the 40-item special pretest sections and sample sizes for each group studied are presented in Table 1. Spiraling of the forms randomized their presentation, and no differences in the ability of the groups across the samples taking the four forms were expected or observed from mean verbal scores; the groups were judged to be essentially parallel across all four forms. Difficulty estimates for the four forms indicated that, for the most part, the four forms were parallel in difficulty. Form A appeared slightly more difficult and Form C slightly easier for most groups studied but, in general, there was a close correspondence among means and standard deviations for the difficulty estimates across all four forms.

Scatterplots of difficulty and discrimination indices (p-values and R-Biserials) are presented in Figures 1 to 4. Figures 1 and 3 present the p-values and Figures 2 and 4 the R-Biserials between Forms A and B and Forms C and D, respectively. These figures show that for those items where the indices follow the diagonal line, the difficulty and discrimination indices remained parallel across forms. More outliers are noted on the difficulty plots than on the discrimination plots. There are about six items per pair of forms with difficulty differences greater than 15 percent. Most of these items were revised by changing words that differed in difficulty, thus affecting the difficulty of the total item. This shift in difficulty was not totally unexpected because, although an effort was made to maintain the relative difficulty of parallel items, previous studies have shown that changes of one word can alter the difficulty of the item (Schmitt, Curley, Bleistein, and Dorans 1988; Bleistein, Schmitt, and Curley 1990).

Tables 2 to 8 present the statistical results associated with all of the different versions of the items studied in this investigation. The tables are organized with the original reprinted version of each question always appearing first (in the left column), regardless of which of the four forms that version may have appeared in. Look, for instance, at Table 2, item 11. The wording of the question

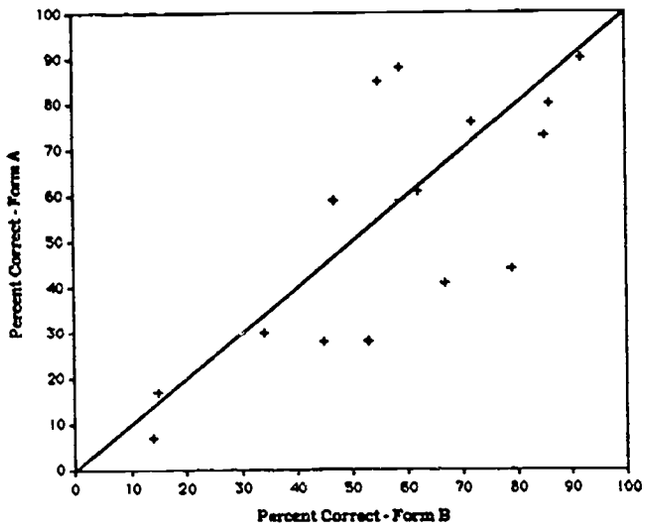

FIGURE 1. Scatterplot of difficulty estimates between pairs of items: Forms A and B.

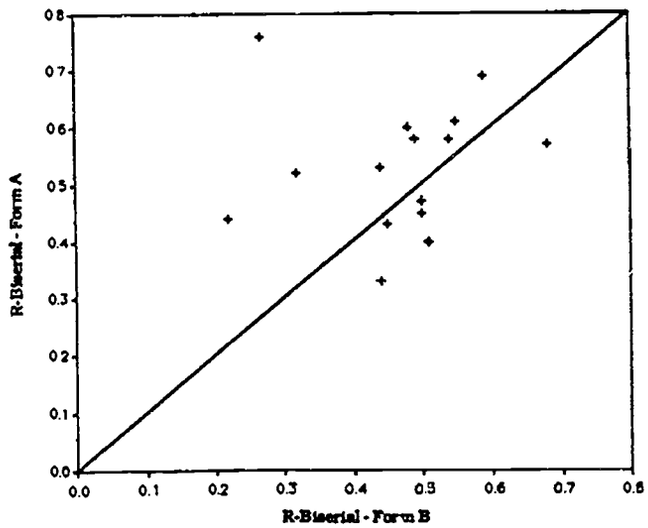

FIGURE 2. Scatterplot of discrimination estimates between pairs of items: Forms A and B.

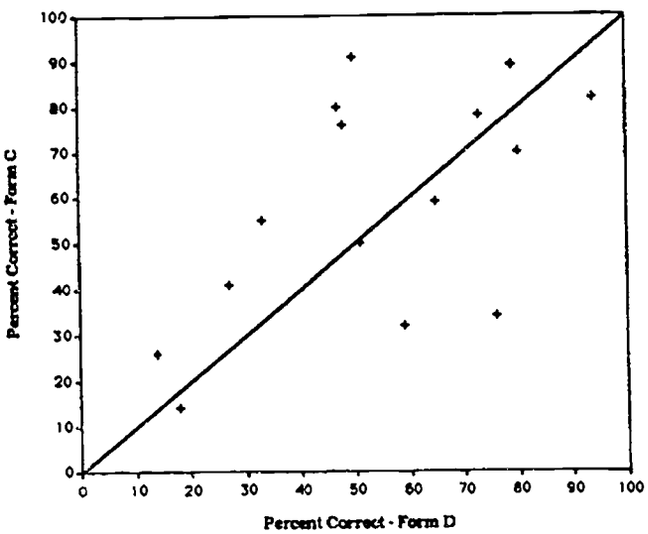

FIGURE 3. Scatterplot of difficulty estimates between pairs of items: Forms C and D.

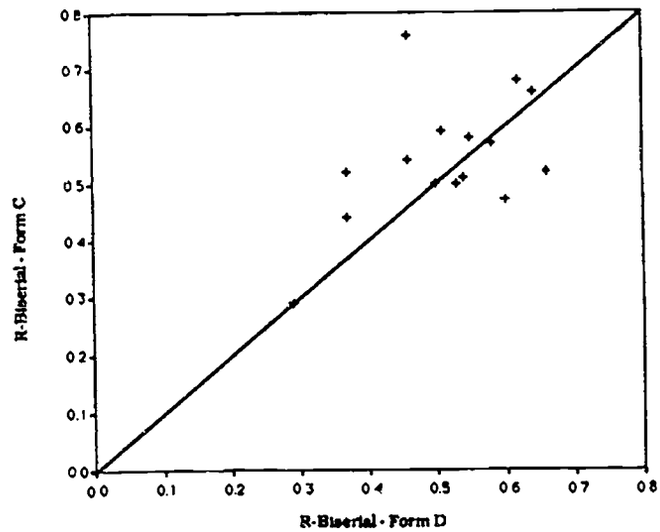

FIGURE 4. Scatterplot of discrimination estimates between pairs of items: Forms C and D.

as it was initially pretested and as it was reprinted for this study (with a key of "curb..predators") appears first, along with the new data. The data indicate that the item was classified as negative "C" for female examinees using the MH metric (–2.15); for Hispanic, black, and Asian American examinees, the item was classified "A." This version of the question appeared in Form A and was answered correctly by 73 percent of the total population; the R-Biserial of the item was .69. To the left of each of the five options (A to E) are found the standardized differences between matched groups of examinees (focal minus reference); the standardized differences for those who omitted each item are also included. For example, in the version of item 11 that appeared in Form A, 12 percent fewer females than matched males selected the key (E), and 11 percent more females than matched males selected distractor (A).

To the right of this first version of item 11 is a second version with revisions indicated in boldface (in the key, "curb" has been changed to "lessen" in Form B). In many cases, only one revision was made to the original item, but for this item (as for several others) there were two additional revisions (in two additional forms) that appear directly below the first two versions of item 11: in Form C the key was changed to "curb..enemies," and in Form D the key was changed to "lessen..enemies." Tables 2 to 8 present the items in this study classified by the hypothesized DIF factors.

TABLE 2

Effects of Science Terminology

Item No.	(Form) P R-Bis	MH D-DIF ETS DIF Category DSTD-P% F	H	B	A	Item Text	(Form) P R-Bis	MH D-DIF ETS DIF Category DSTD-P% F	H	B	A	Item Text
11.	(A) .73 .69	-2.15 C	-0.56 A	-0.78 A	-0.58 A	In order to —— the health hazard caused by an increased pigeon population, officials have added to the area's number of peregrine falcons, natural —— of pigeons.	(B) .85 .59	-1.42 B	-0.21 A	-1.04 B	0.38 A	In order to —— the health hazard caused by an increased pigeon population, officials have added to the area's number of peregrine falcons, natural —— of pigeons.
		11	4	6	0	(A) reduce..allies		6	0	8	0	(A) reduce..allies
		1	1	0	2	(B) promote..rivals		1	-1	-1	0	(B) promote..rivals
		0	0	0	0	(C) starve..prey		0	1	0	1	(C) starve..prey
		0	-2	0	1	(D) counter..protectors		0	0	0	-1	(D) counter..protectors
		-12	-4	-6	-3	*(E) curb..predators		-6	-1	-8	0	*(E) lessen..predators
		0	1	1	0	(OMITS)		0	1	1	0	(OMITS)
11.	(C) .70 .68	-1.94 C	-1.12 B	-0.56 A	-0.50 A	In order to —— the health hazard caused by an increased pigeon population, officials have added to the area's number of peregrine falcons, natural —— of pigeons.	(D) .80 .62	-1.79 C	-0.15 A	-0.92 A	-0.20 A	In order to —— the health hazard caused by an increased pigeon population, officials have added to the area's number of peregrine falcons, natural —— of pigeons.
		10	6	5	2	(A) reduce..allies		9	2	4	0	(A) reduce..allies
		1	2	1	2	(B) promote..rivals		0	0	1	1	(B) promote..rivals
		-1	3	-1	0	(C) starve..prey		0	0	2	0	(C) starve..prey
		1	-1	0	-1	(D) counter..protectors		0	0	1	0	(D) counter..protectors
		-12	-8	-4	-3	*(E) curb..enemies		-9	-1	-8	-1	*(E) lessen..enemies
		0	-1	0	0	(OMITS)		0	-1	0	0	(OMITS)
12.	(B) .59 .68	-1.75 C	-0.04 A	-0.94 A	-0.83 A	Although cacti are not —— Hawaii, several species have been introduced there and are flourishing.	(A) .88 .57	-1.30 B	-1.18 B	-1.09 B	-1.91 C	Although cacti are not —— Hawaii, several species have been introduced there and are flourishing.
		10	6	7	8	(A) adaptable to		4	7	6	3	(A) adaptable to
		-13	-1	-7	-6	*(B) indigenous to		-5	-7	-7	-6	*(B) native to
		0	0	2	0	(C) excluded from		0	0	2	2	(C) excluded from
		3	-5	-2	-2	(D) compatible with		1	0	0	0	(D) compatible with
		0	0	0	-1	(E) limited to		0	1	-1	0	(E) limited to
		0	0	0	0	(OMITS)		0	0	1	0	(OMITS)
21.	(A) .41 .47	-2.27 C	-0.33 A	-0.60 A	0.53 A	CHIMPANZEE:PRIMATE::	(B) .67 .50	-1.01 B	-0.23 A	-0.38 A	-0.32 A	CHIMPANZEE:PRIMATE::
		0	4	3	-1	(A) baboon:gorilla		1	1	-1	1	(A) baboon:gorilla
		1	2	5	1	(B) cat:kitten		1	4	2	0	(B) cat:kitten
		2	0	0	-3	(C) cocoon:larva		4	-2	-1	1	(C) cocoon:larva
		-20	-2	-5	5	*(D) squirrel:rodent		-8	-2	-3	-3	*(D) squirrel:rodent
		15	-2	-4	-1	(E) fish:amphibian		1	1	0	-1	(E) fish:gill
		2	-1	1	-2	(OMITS)		2	-2	2	1	(OMITS)
21.	(D) .48 .46	-1.51 C	-0.23 A	-0.50 A	-0.11 A	CHIMPANZEE:PRIMATE::	(C) .76 .54	-0.79 A	-0.22 A	-0.81 A	0.84 A	CHIMPANZEE:PRIMATE::
		-1	-1	1	1	(A) baboon:gorilla		0	1	1	0	(A) baboon:gorilla
		0	1	5	1	(B) cat:kitten		2	0	6	-1	(B) cat:kitten
		0	3	1	1	(C) cocoon:larva		1	4	1	-2	(C) cocoon:larva
		-14	-2	-5	-1	*(D) mouse:rodent		-5	-2	-7	4	*(D) frog:amphibian
		12	2	-4	0	(E) fish:amphibian		0	1	0	-1	(E) squirrel:reptile
		2	-3	3	-2	(OMITS)		2	-4	0	0	(OMITS)
24	(C) .34 .45	-2.28 C	0.87 A	0.17 A	0.67 A	VORTEX:WATER::	(D) .76 .37	-0.83 A	-0.81 A	-0.58 A	0.38 A	WHIRLPOOL:WATER::
		1	-2	0	1	(A) volcano:crust		1	1	1	0	(A) volcano:crust
		2	-1	3	-2	(B) river:delta		1	0	3	0	(B) river:delta
		-19	7	1	6	*(C) tornado:air		-6	-7	-5	2	*(C) tornado:air
		3	2	2	3	(D) geyser:steam		3	4	0	-2	(D) geyser:steam
		-1	2	1	-3	(E) earthquake:fault		1	1	1	0	(E) earthquake:fault
		14	-9	-8	-4	(OMITS)		0	1	2	0	(OMITS)

MH D-DIF: Mantel-Haenszel Index of Delta Differences (focal minus reference)
ETS DIF Category: A represents negligible DIF, B represents slight to moderate DIF, and C represents moderate to large DIF.
DSTD-P%: Standardization Index of Proportion Correct Differences (focal minus reference)
F: matched female/male comparison H: matched Hispanic/white comparison
B: matched black/white comparison A: matched Asian American/white comparison
*Indicates correct answer.
Item revisions are indicated by boldface.

Two Initial Observations

Before turning to the results of this investigation that speak to the primary purposes of the study, two related observations based on the data warrant some initial consideration. First is the issue of variation in DIF data for identically repeated items between initial pretesting and subsequent reprinting for this study. Of course a certain amount of "noise" should always be expected in such data simply because of differences in the samples and the contexts (i.e., surrounding items) in which the repeated items appear. It should not be surprising, for example, to see more variation in the DIF data of identically reprinted items for minority examinees than for female examinees, given that minority sample sizes are generally much smaller than other sample sizes. (In fact, the standard error of the MH delta statistic is about .50 for minority groups on SAT-V, while for the male/female comparison the standard error is about .15.) Of the 23 items reprinted for this study, there were 6 for which the ETS DIF category shifted from "C" to "B" or "A" for identically reprinted items.

Table 9 shows MH D-DIF values for the six items for which such shifts in the ETS DIF categories occurred. There were also some items with "B" to "A" or "A" to "B" shifts for some groups, but these were not included in the table because they are not relevant to this study. All but two of the six items (Form C, item 13, and Form D, item 25) showed shifts in values that were within two standard errors of the MH D-DIF statistic. (Differences of more than two standard errors are expected to occur less than 5 percent of the time by chance alone.) Note that two of the identically reprinted items (Form D, items 16 and 23) each shifted two categories for one of the focal groups—from "C" at initial pretesting to "A" in this study—even though both shifts were within sampling error. Note also that another of the items (Form D, item 25) actually shifted two categories and changed sign, from a positive "C" for Hispanics at initial pretesting to a negative "A" for Hispanics in this study; this shift in value was beyond that which would be expected given sampling error.

The second general observation from these data not directly related to the primary purposes of the study in-

TABLE 3

Effects of Industrial Arts Terminology

Item No.	(Form) P R-Bis	MH D-DIF ETS DIF Category DSTD-P% F	H	B	A	Item Text	(Form) P R-Bis	MH D-DIF ETS DIF Category DSTD-P% F	H	B	A	Item Text
19.	(B)	-3.25	-2.03	-1.49	-0.86	RIVET:METAL::	(A)	0.37	-0.99	-0.98	-0.13	PIN:CLOTH::
	.55	C	C	B	A		.85	A	A	A	A	
	.45	3	3	2	4	(A) needle:thimble	.43	0	2	4	0	(A) needle:thimble
		1	2	-1	-2	(B) cork:bottle		-1	3	0	-1	(B) cork:bottle
		5	8	4	2	(C) nail:hammer		0	0	1	0	(C) nail:hammer
		-28	-18	-13	-8	*(D) staple:paper		2	-6	-7	0	*(D) staple:paper
		3	1	2	-2	(E) rope:swing		-1	1	1	0	(E) rope:swing
		16	4	7	5	(OMITS)		0	0	0	0	(OMITS)
23.	(D)	-1.93	-0.72	-0.83	-0.84	BIT:DRILL::	(C)	-1.50	-0.90	-1.24	-0.66	BIT:DRILL::
	.33	C	A	A	A		.55	B	A	B	A	
	.54	0	-5	-1	3	(A) wax:cravon	.51	1	0	0	0	(A) wax:cravon
		-1	1	2	2	(B) impression:stylus		1	1	3	2	(B) impression:stylus
		9	-1	-2	-2	(C) handle:brush		5	-1	-2	-2	(C) handle:brush
		-15	-5	-5	-7	*(D) point:awl		-13	-8	-11	-5	*(D) point:spear
		-2	-2	0	-2	(E) needle:thread		0	4	2	-4	(E) needle:thread
		9	12	5	7	(OMITS)		-7	3	8	9	(OMITS)
23.	(A)	-1.96	-2.20	-1.67	-0.21	PRONGS:PITCHFORK::	(B)	-1.03	-2.25	-1.18	-2.42	PRONGS:PITCHFORK::
	.44	C	C	C	A		.79	B	C	B	C	
	.58	2	0	1	-2	(A) wax:cravon	.49	2	2	-1	1	(A) wax:cravon
		0	3	0	3	(B) impression:stylus		-1	2	1	2	(B) impression:stylus
		13	3	0	-7	(C) handle:brush		5	3	-1	4	(C) handle:brush
		-16	-16	-11	-1	*(D) point:awl		-6	-17	-10	-15	*(D) point:spear
		0	6	3	0	(E) needle:thread		0	3	3	2	(E) needle:thread
		1	5	6	7	(OMITS)		1	-7	8	6	(OMITS)

MH D-DIF: Mantel-Haenszel Index of Delta Differences (focal minus reference)
ETS DIF Category: A represents negligible DIF, B represents slight to moderate DIF, and C represents moderate to large DIF
DSTD-P%: Standardization Index of Proportion Correct Differences (focal minus reference)
F: matched female/male comparison H: matched Hispanic/white comparison
B: matched black/white comparison A: matched Asian American/white comparison
*Indicates correct answer.
Item revisions are indicated by boldface.

TABLE 4

Effects of Military Terminology

Item No.	(Form) P R-Bis	MH D-DIF ETS DIF Category DSTD-P% F	H	B	A	Item Text	(Form) P R-Bis	MH D-DIF ETS DIF Category DSTD-P% F	H	B	A	Item Text
17.	(C)	-3.69	-1.09	-1.23	-0.81	CONVOY:SHIPS::	(D)	0.56	0.16	0.21	0.39	TROUPE:DANCERS::
	.78	C	B	B	A		.73	A	A	A	A	
	.47	-21	-9	-10	-4	*(A) flock:birds	.60	4	1	2	2	*(A) flock:birds
		4	5	5	1	(B) ferry:passengers		0	2	4	2	(B) ferry:passengers
		4	3	4	1	(C) barn:horses		-2	-1	0	0	(C) barn:horses
		5	0	2	-1	(D) dealership:cars		0	0	-2	-2	(D) dealership:cars
		3	2	-1	-1	(E) highway:trucks		-1	-1	-2	-1	(E) highway:trucks
		6	0	1	4	(OMITS)		-1	0	-3	-1	(OMITS)
18.	(A)	-2.84	-0.40	-0.37	0.41	DETONATE:EXPLOSION::	(B)	-0.32	-0.64	-0.63	-0.41	PROVOKE:REACTION::
	.76	C	A	A	A		.72	A	A	A	A	
	.33	1	1	0	1	(A) collide:momentum	.44	0	1	1	1	(A) collide:momentum
		5	1	0	-3	(B) decipher:code		1	0	1	-1	(B) decipher:code
		3	2	2	-1	(C) energize:stimulant		1	5	5	3	(C) energize:stimulant
		1	1	1	0	(D) strike:ore		0	1	0	0	(D) strike:ore
		-19	-4	-2	3	*(E) ignite:fire		-3	-6	-7	-3	*(E) ignite:fire
		9	-1	-1	0	(OMITS)		0	-1	1	0	(OMITS)
19.	(D)	-2.75	-0.02	-0.83	-0.52	AMMUNITION:CARTRIDGE	(C)	-2.34	-1.14	-0.92	-0.36	MONEY:WALLET::
	.65	C	A	A	A	BELT::	.59	C	B	A	A	
	.50	3	-1	3	2	(A) rifle:trigger	.50	1	1	0	-2	(A) rifle:trigger
		5	3	4	3	(B) dart:spear		3	5	5	1	(B) dart:spear
		-21	0	-9	-4	*(C) arrow:quiver		-20	-10	-9	-3	*(C) arrow:quiver
		0	1	1	0	(D) golf:course		2	5	2	2	(D) golf:course
		8	-1	-1	0	(E) football:goalpost		12	-2	1	2	(E) football:goalpost
		5	-3	3	-1	(OMITS)		1	1	2	0	(OMITS)
20.	(C)	-2.32	-0.17	-0.46	0.41	MUTINY:CAPTAIN::	(D)	-2.00	0.33	-0.08	0.61	MUTINY:CAPTAIN::
	.50	C	A	A	A		.51	C	A	A	A	
	.58	1	-1	2	-2	(A) theft:police	.55	3	2	2	-3	(A) theft:police
		1	1	-1	-1	(B) riot:crowd		2	2	-1	0	(B) riot:crowd
		3	-1	4	-3	(C) plagiarism:author		3	-3	-1	-4	(C) plagiarism:author
		-19	2	-4	3	*(D) strike:employer		*-17	3	-1	5	*(D) strike:employer
		6	4	1	-1	(E) war:general		1	1	4	1	(E) recipe:chef
		9	-5	-1	3	(OMITS)		8	-4	-2	1	(OMITS)
20.	(A)	-0.73	-0.62	0.43	0.64	REBELLION:AUTHORITY::	(B)	-0.84	-0.72	0.18	0.55	REBELLION:AUTHORITY::
	.61	A	A	A	A		.62	A	A	A	A	
	.58	-1	1	-7	-4	(A) theft:police	.54	4	3	-6	-3	(A) theft:police
		1	2	2	0	(B) riot:crowd		2	1	0	3	(B) riot:crowd
		2	-2	0	1	(C) plagiarism:author		1	2	2	-2	(C) plagiarism:author
		-6	-5	3	4	*(D) strike:employer		-7	-6	0	3	*(D) strike:employer
		2	4	1	-2	(E) war:general		0	1	0	0	(E) recipe:chef
		1	0	1	0	(OMITS)		0	-2	4	-1	(OMITS)

volves the easiest and most difficult items in the SAT-V (extremes in the difficulty continuum). See, for instance, Table 8, item 12, Forms D and C. Since both versions were classified as negative "C" for females, it would appear that the revision of this item (changing the key from "tapping" to "utilizing") did not succeed in eliminating the DIF. Yet the differences in proportion correct between matched groups of males and females—which are reported using Standardization rather than Mantel-Haenszel—indicate a shift from -23 to -5. That is, the version of item 12 in Form C shows a small DSTD p-metric value (a 5 percent difference in performance between matched males and females), yet (using the MH delta metric) it was still categorized as "C." Note, however, that the revision to this sentence completion item changed it from a middle difficulty item (50 percent correct) to a very easy item (better than 90 percent correct).

The same sort of phenomenon can be observed at the other end of the difficulty scale. See Table 4, item 25, Form A. This item was classified as negative "C" for females (using the MH delta metric) despite the fact that there is only a 4 percent difference in performance on the item between matched groups of males and females (using the DSTD p metric). Note again, however, the extreme overall difficulty of the item: only 7 percent of the total population answered this question correctly. Data on the revised version of the item in Form B (with stem and key switched) show that the ETS DIF category changed from "C" to "B" for females, yet the difference in performance between matched groups of males and

TABLE 4 *(continued)*

Effects of Military Terminology

Item No.	(Form) P R-Bis	MH D-DIF ETS DIF Category DSTD-P% F	H	B	A	Item Text	(Form) P R-Bis	MH D-DIF ETS DIF Category DSTD-P% F	H	B	A	Item Text
22.	(B)	−4.31	−1.05	−1.22	−1.07	COCKPIT:PILOT::	(A)	−3.31	−1.01	−0.89	−0.23	COCKPIT:PILOT::
	.47	C	B	B	B		.59	C	B	A	A	
	.44	−37	−9	−10	−10	*(A) turret:gunner	.53	−26	−9	−8	−2	*(A) turret:gunner
		1	2	2	2	(B) somersault:acrobat		1	1	3	1	(B) somersault:acrobat
		25	−3	−5	0	(C) berth:sailor		1	3	−1	0	(C) uniform:fire fighter
		0	2	−1	−2	(D) baton:conductor		1	1	3	0	(D) baton:drum major
		10	6	8	6	(E) sidewalk:pedestrian		22	5	−1	−1	(E) sidewalk:pedestrian
		1	2	6	3	(OMITS)		1	−1	4	2	(OMITS)
22.	(C)	−0.54	−1.99	−1.92	−2.15	COCKPIT:PILOT::	(D)	−0.07	−0.99	−0.62	−0.87	STALL:VENDOR::
	.89	A	C	C	C		.79	A	A	A	A	
	.57	−2	−10	−12	−7	*(A) booth:toll collector	.58	0	−6	−4	−4	*(A) booth:toll collector
		0	3	1	1	(B) somersault:acrobat		0	3	3	1	(B) somersault:acrobat
		0	1	0	0	(C) uniform:fire fighter		−1	−1	1	0	(C) uniform:fire fighter
		0	1	2	0	(D) baton:drum major		0	0	1	0	(D) baton:drum major
		1	3	6	4	(E) sidewalk:pedestrian		2	5	2	3	(E) sidewalk:pedestrian
		1	2	4	2	(OMITS)		0	−1	−3	0	(OMITS)
25.	(A)	−1.71	0.50	0.75	0.45	MERCENARY:WARFARE::	(B)	−1.29	−0.02	−0.61	−0.92	HACK:WRITING::
	.07	C	A	A	A		.14	B	A	A	A	
	.44	0	−2	3	0	(A) truant:school	.22	1	0	1	3	(A) truant:school
		−1	2	2	0	(B) thief:property		−4	−2	2	−1	(B) thief:property
		−4	−1	1	2	*(C) hack:writing		−6	0	−3	−5	*(C) mercenary:warfare
		−12	9	1	2	(D) criminal:felony		1	2	4	2	(D) criminal:felony
		6	0	−3	−3	(E) defendant:accusation		1	6	4	1	(E) defendant:accusation
		11	−8	−5	−1	(OMITS)		7	−6	−8	−1	(OMITS)

MH D-DIF: Mantel-Haenszel Index of Delta Differences (focal minus reference)
ETS DIF Category: A represents negligible DIF, B represents slight to moderate DIF, and C represents moderate to large DIF.
DSTD-P%: Standardization Index of Proportion Correct Differences (focal minus reference)
F: matched female/male comparison H: matched Hispanic/white comparison
B: matched black/white comparison A: matched Asian American/white comparison
*Indicates correct answer.
Item revisions are indicated by boldface.

females actually increased slightly (6 percent). On the revised version of this item, however, twice as many examinees answered correctly overall (14 percent of the total population).

A possible explanation for these apparent DIF anomalies among the very easy and difficult SAT-V items is related to the nature of the two different metrics being used. The Mantel-Haenszel index represents odds ratios converted to the ETS delta scale, a scale that is unbounded at the two ends. The Standardization index, on the other hand, represents differences in proportions correct on a scale (0 to 100) that is bounded at the top and bottom (Dorans and Holland 1992). Thus the two sets of statistics often behave differently with the easiest and most difficult SAT-V items.

Success Rate in Reducing Differential Item Functioning

One of the primary purposes of this study was to determine how successfully SAT-V items with elevated levels of DIF—particularly those items for which the DIF seemed to be related to a factor under study—could be revised in order to reduce or eliminate differential functioning. In this way, it could be determined whether or not similar efforts at revision and re-pretesting would be worthwhile in the future.

This investigation began with 23 items but, as discussed above, some of them did not demonstrate C DIF after being reprinted and re-pretested for this study. Of the 18 items that actually fell into category "C" when they were reprinted, 12 items (67 percent) successfully shifted from C DIF to B DIF or A DIF after being pretested with the revisions. All 12 of the items that changed from C DIF showed reductions in MH D-DIF values outside those expected within sampling error. (All five items that did not fall into category "C" for any group when reprinted for this study showed small to moderate reductions in MH and/or DSTD values after the revisions.) Of the 12 items that shifted from C DIF, 9 of them shifted from negative C DIF and 3 from positive C DIF; since 15 of the 18 actual C DIF items were negative and 3 of the 18 were positive, there was a 60 percent success rate in eliminating negative C DIF (i.e., DIF not

favoring focal groups) and a 100 percent success rate in eliminating positive C DIF (i.e., DIF favoring focal groups).

A closer look at the six items that did not shift from C DIF reveals that, in two cases, the revisions did indeed eliminate the C DIF for the group originally targeted, but a different group ended up with C DIF (Table 2, item 12, and Table 3, item 23). In another case, the revision successfully eliminated the C DIF but the R-Biserial ended up below .30, which meant the revised item could not be included in the pool (Table 4, item 25). In a fourth case, the MH value shifted dramatically in the intended direction but, as discussed above, the revised item became very easy for the total population and was still classified "C" using the MH delta metric (Table 8, item 12). Thus, in only two of the total of 23 items (Table 2, item 11, and Table 4, item 19) did no appreciable reduction in DIF occur for the targeted groups(s) after the revisions were made.

It must be mentioned, however, that 6 of the 12 items for which C DIF was successfully eliminated also became substantially easier for the total population, i.e., percent correct increased by 25 percent or more (Table 2, items 21 and 24; Table 3, item 19; Table 4, item 22; Table 6, items 14 and 18). Another 2 of the 12 items shifted content classification as a result of revisions to the stems (Table 4, items 17 and 18). So in 8 out of 12 cases, the successfully revised items were significantly changed from the original versions either in content or statistics. For SAT-V items, elimination of C DIF often seemed to change some basic characteristic(s) of the item, yet the underlying reasoning skill being tested remained essentially the same in most cases. Assuming that item pools are large enough to allow assemblers to continue to meet test specifications, such shifts in content or statistics seem less important than the fact that the items no longer show elevated levels of DIF and thus can be considered for inclusion in operational forms of the test.

Factors Related to Differential Item Functioning

Because only a limited number of items were pretested for each of the seven factors studied, and because the revisions of some of the items changed the degree of difficulty of the item considerably, conclusions about the relationship between the DIF factors studied and the observed DIF values must be made with caution.

Effects of Science Terminology

Technical (specialized) science material and substantive contexts drawn from science have been found to affect negatively the performance of female examinees on the SAT-V (Lawrence, Curley, and McHale 1988; Lawrence and Curley 1989; Scheuneman and Gerritz 1990). A look at Table 2 reveals that C DIF for females was eliminated after the revisions were made in three of the four science items included in this study; the other item (11) showed some reduction in the MH value in Form B (–1.42), but the further revisions in Forms C and D showed a return of negative C DIF for females. In item 12 the change in the key from "indigenous" to "native" eliminated the C DIF for females but introduced C DIF for Asian American examinees; the item also became much easier overall (88% correct) because of the revision. Items 21 and 24 became markedly easier, too, after the revisions were made. Note in item 21 that females were attracted differentially whenever "fish:amphibian" was used as a wrong answer choice (Forms A and D), but the version in Form C worked very well with "frog:amphibian" as the key. In item 24, females differentially omitted the item when "VORTEX" was in the stem but not when "WHIRLPOOL" was in the stem.

Effects of Industrial Arts Terminology

The revisions made in both of the items shown in Table 3 significantly lowered the elevated levels of negative DIF for female examinees but also made both of the items considerably easier for the total population. In item 19, after changing "RIVET:METAL" to "PIN:CLOTH," the differential percentage of matched females who omitted the item was reduced from 16 to zero. Also in item 19, the level of negative DIF for Hispanic and black examinees (as well as for females) was greatly reduced. In item 23, high levels of DIF against matched females were reduced only when both the stem and key were revised but, with the introduction of the new stem ("PRONGS:PITCHFORK"), larger amounts of negative DIF for Hispanic and black examinees appeared (Form A). Then, with the addition of the new key ("point:spear"), negative C DIF for Hispanic and Asian American examinees was observed (Form B).

Effects of Military Terminology

Table 4 reveals that negative C DIF for females was successfully eliminated in four out of the six items when the analogy stems "CONVOY:SHIPS," "DETONATE:EXPLOSION," "MUTINY:CAPTAIN," and "COCKPIT:PILOT" were changed (respectively) to "TROUPE:DANCERS," "PROVOKE:REACTION," "REBELLION:AUTHORITY," and "STALL:VENDOR." The revisions in the stems of items 17 and 18 also changed the content categories from "Practical Affairs/Social Sciences" to "Humanities/Human Relations," but the overall difficulty levels remained approximately the same. The revision of the stem of item 19 did not eliminate the nega-

TABLE 5

Effects of Contexts Portraying Aggression or Conflict

Item No.	(Form) P R-Bis	MH D-DIF ETS DIF Category DSTD-P% F	H	B	A	Item Text
13.	(C) .41 .76	-1.35 B	0.25 A	0.12 A	0.61 A	These ominous developments suggest that the political conflict in that country has entered a new and more —— phase.
		2	4	0	−1	(A) moderate
		0	1	3	−1	(B) legitimate
		7	−3	−3	−3	(C) productive
		−9	1	1	4	*(D) perilous
		−1	1	0	0	(E) inconsequential
		1	−4	−0	1	(OMITS)
13.	(A) .28 .76	−0.96 A	−0.25 A	−0.35 A	−0.05 A	These ominous developments suggest that the social climate in that country has entered a new and more —— phase.
		0	0	6	−2	(A) moderate
		−1	2	1	2	(B) legitimate
		5	2	−2	0	(C) productive
		−5	−1	−1	0	*(D) perilous
		0	−1	0	1	(E) inconsequential
		1	−2	−3	0	(OMITS)
15.	(B) .15 .55	−1.64 C	−0.30 A	−0.33 A	0.44 A	Heretofore —— for his emphasis on defensive strategies, the general was —— when doctrines emphasizing aggression were discredited.
		−1	2	−1	0	(A) criticized..discharged
		1	−3	0	4	(B) parodied..ostracized
		7	2	2	−6	(C) supported..disappointed
		−7	−1	0	2	*(D) spurned..vindicated
		1	0	2	−2	(E) praised..disregarded
		0	1	−3	1	(OMITS)
15.	(D) .18 .51	−1.56 C	−0.28 A	−0.21 A	0.04 A	Heretofore —— for his emphasis on defensive strategies, the general was —— when doctrines emphasizing aggression were discredited.
		−1	−1	0	1	(A) criticized..discharged
		1	1	0	−1	(B) parodied..ostracized
		9	4	7	−4	(C) supported..disappointed
		−8	−2	0	0	*(D) chastised..vindicated
		0	3	−2	0	(E) praised..disregarded
		−1	−5	−4	4	(OMITS)

Item No.	(Form) P R-Bis	MH D-DIF ETS DIF Category DSTD-P% F	H	B	A	Item Text
	(D) .27 .46	−0.99 A	0.42 A	0.96 A	0.74 A	These inauspicious developments suggest that the political conflict in that country has entered a new and more —— phase.
		−2	−5	−2	−2	(A) moderate
		−1	−2	1	−3	(B) legitimate
		7	3	1	−3	(C) productive
		−7	3	4	6	*(D) perilous
		0	0	−1	1	(E) inconsequential
		4	1	−3	1	(OMITS)
	(B) .53 .27	0.16 A	-0.38 A	-0.05 A	0.33 A	These auspicious developments suggest that the political climate in that country has entered a new and more —— phase.
		−4	−1	−1	−3	(A) hazardous
		−1	4	0	2	(B) illegitimate
		0	0	4	1	(C) unproductive
		2	−4	−1	3	*(D) promising
		1	2	0	−1	(E) inconsequential
		3	−1	−2	−2	(OMITS)
	(A) .17 .61	−2.33 C	0.89 A	0.15 A	0.36 A	In the past the general had been —— for his emphasis on defensive strategies, but he was —— when doctrines emphasizing aggression were discredited.
		−3	2	−2	−1	(A) criticized..discharged
		1	2	0	0	(B) parodied..ostracized
		5	−9	2	−5	(C) supported..disappointed
		−11	3	1	2	*(D) spurned..vindicated
		8	3	1	2	(E) praised..disregarded
		0	0	−1	2	(OMITS)
	(C) .14 .59	−1.27 B	0.26 A	0.25 A	−0.04 A	Heretofore —— for her emphasis on conservation, the economist was —— when doctrines emphasizing consumption were discredited.
		−2	5	−1	−3	(A) criticized..discharged
		0	−1	3	3	(B) parodied..ostracized
		7	1	2	0	(C) supported..disappointed
		−5	1	1	0	*(D) spurned..vindicated
		2	−1	−2	−6	(E) praised..disregarded
		−3	−5	−3	6	(OMITS)

MH D-DIF: Mantel-Haenszel Index of Delta Differences (focal minus reference)
ETS DIF Category: A represents negligible DIF, B represents slight to moderate DIF, and C represents moderate to large DIF.
DSTD-P%: Standardization Index of Proportion Correct Differences (focal minus reference)
F: matched female/male comparison H: matched Hispanic/white comparison
B: matched black/white comparison A: matched Asian American/white comparison
*Indicates correct answer.
Item revisions are indicated by boldface.

tive C DIF for female examinees, although a further revision of the key ("arrow:quiver") might have helped to produce the intended effect.

Item 20 reveals that changing only a distractor ("war:general" to "recipe:chef" in Form D) did not eliminate the negative C DIF for females; rather, "mutiny" was the term that females seemed less familiar with than the matched groups of males. Item 22 is interesting in that the terms in the key, "turret:gunner," seemed more differentially difficult for females than the terms in the stem, "COCKPIT:PILOT" (see the version in Form C). With the change to a new key in Form C ("booth:toll collector"), however, negative C DIF was present for all three minority groups. The C DIF was eliminated entirely in Form D when the stem was revised as well to "STALL:VENDOR." Item 25 was discussed earlier; the R-Biserial of .22 in Form B makes this version unacceptable for use in the pool even though the MH value was reduced.

Contexts Portraying Aggression/Conflict

Items suggesting aggression or conflict as well as items with a strongly negative, possibly upsetting tone have

TABLE 6

Effects of Special Interest Terminology

Item No.	(Form) P R-Bis	MH D-DIF ETS DIF Category DSTD-P% F	H	B	A	Item Text	(Form) P R-Bis	MH D-DIF ETS DIF Category DSTD-P% F	H	B	A	Item Text
14.	(C)	2.15	−0.41	−0.86	0.31	It is its ——, its rhythmic energy	(D)	1.19	0.61	0.25	0.98	It is its ——, its rhythmic energy
	.32	C	A	A	A	and expansive vivacity, that	.59	B	A	A	A	and expansive vivacity, that
	.66					makes jazz so typically American.	.64					makes jazz so typically American.
		14	−2	−4	2	*(A) verve		9	4	2	7	*(A) vitality
		−4	2	−1	−3	(B) paucity		−2	−3	−1	−2	(B) paucity
		0	9	5	0	(C) formality		0	−1	3	1	(C) formality
		−3	0	3	1	(D) quiescence		−1	3	1	−3	(D) quiescence
		−2	−1	−3	−1	(E) derivativeness		−2	−1	−1	−2	(E) derivativeness
		−6	−8	−1	1	(OMITS)		−3	−2	−3	−2	(OMITS)
24.	(A)	1.53	1.01	0.97	−0.08	DOTE:FONDNESS::	(B)	1.06	1.05	0.62	0.53	ABHOR:DISTASTE::
	.30	C	B	A	A		.34	B	B	A	A	
	.45	0	3	0	1	(A) improvise:practice	.50	1	3	0	−1	(A) improvise:practice
		−4	−1	2	1	(B) attract:repulsion		−2	−2	1	−2	(B) attract:repulsion
		12	7	7	0	*(C) pamper:indulgence		8	8	5	4	*(C) pamper:indulgence
		−2	−1	1	0	(D) unnerve:composure		−1	1	0	4	(D) unnerve:composure
		3	0	1	−1	(E) supervise:regulation		0	−2	2	1	(E) supervise:regulation
		−10	−8	−10	−1	(OMITS)		−6	−7	−6	−8	(OMITS)
18.	(D)	0.81	−0.88	2.08	0.36	PLAIT:HAIR::	(C)	−0.28	−0.72	−0.39	−0.34	BRAID:HAIR::
	.47	A	A	C	A		.80	A	A	A	A	
	.29	0	1	1	1	(A) knead:bread	.29	2	−1	0	1	(A) knead:bread
		8	−8	21	3	*(B) weave:yarn		−2	−6	−4	−2	*(B) weave:yarn
		0	3	−3	−1	(C) cut:cloth		0	2	1	1	(C) cut:cloth
		−2	4	2	−1	(D) fold:paper		−1	1	−1	0	(D) fold:paper
		1	2	−2	1	(E) frame:picture		1	4	3	0	(E) frame:picture
		−6	−2	−19	−2	(OMITS)		0	0	1	1	(OMITS)

MH D-DIF: Mantel-Haenszel Index of Delta Differences (focal minus reference)
ETS DIF Category: A represents negligible DIF, B represents slight to moderate DIF, and C represents moderate to large DIF.
DSTD-P%: Standardization Index of Proportion Correct Differences (focal minus reference)
F: matched female/male comparison H: matched Hispanic/white comparison
B: matched black/white comparison A: matched Asian American/white comparison
*Indicates correct answer.
Item revisions are indicated by boldface.

TABLE 7

Effects of Cognates

Item No.	(Form) P R-Bis	MH D-DIF ETS DIF Category DSTD-P% F	H	B	A	Item Text	(Form) P R-Bis	MH D-DIF ETS DIF Category DSTD-P% F	H	B	A	Item Text
14.	(A)	−1.33	−1.19	−0.52	−0.78	Although scholars often wrestle	(B)	−0.18	−0.20	0.39	−0.55	Although scholars often wrestle
	.28	B	B	A	A	with how to —— the impact of	.45	A	A	A	A	with how to —— the impact of
	.60					various influences in an author's	.48					various influences in an author's
						life on that author's work, they						life on that author's work, they
						sometimes neglect to —— the						sometimes neglect to —— the
						effect of such forces on their own						effect of such forces on their own
						writings.						writings.
		−1	0	1	3	(A) quantify..expunge		−1	−1	1	2	(A) quantify..expunge
		1	1	−1	1	(B) surmise..censor		−1	2	−1	1	(B) surmise..censor
		8	8	1	0	(C) evaluate..amplify		5	−1	−2	1	(C) evaluate..amplify
		−9	−6	−2	−5	*(D) gauge..scrutinize		−2	−2	3	−5	*(D) measure..scrutinize
		1	0	3	−3	(E) disguise..amend		−1	1	0	0	(E) disguise..amend
		0	−2	−1	4	(OMITS)		0	−1	−1	0	(OMITS)
25.	(D)	0.14	−0.09	1.01	0.39	DULCET:TONE::	(C)	0.01	−0.21	0.03	0.72	EUPHONIOUS:TONE::
	.14	A	A	B	A		.26	A	A	A	A	
	.37	1	3	1	0	(A) pleased:smile	.52	0	−1	2	−1	(A) pleased:smile
		0	8	3	0	(B) monotonous:voice		5	6	5	−1	(B) monotonous:voice
		0	−2	0	3	(C) insatiable:appetite		−1	4	2	0	(C) insatiable:appetite
		1	5	1	−2	(D) sarcastic:wit		0	4	−1	−2	(D) sarcastic:wit
		0	0	4	2	*(E) delicious:taste		0	−1	0	6	*(E) delicious:taste
		−2	−13	−10	−4	(OMITS)		−3	−12	−8	−2	(OMITS)

MH D-DIF: Mantel-Haenszel Index of Delta Differences (focal minus reference)
ETS DIF Category: A represents negligible DIF, B represents slight to moderate DIF, and C represents moderate to large DIF.
DSTD-P%: Standardization Index of Proportion Correct Differences (focal minus reference)
F: matched female/male comparison H: matched Hispanic/white comparison
B: matched black/white comparison A: matched Asian American/white comparison
*Indicates correct answer.
Item revisions are indicated by boldface.

TABLE 8

Effects of Homographs

Item No.	(Form) P R-Bis	MH D-DIF ETS DIF Category DSTD-P% F	H	B	A	Item Text	(Form) P R-Bis	MH D-DIF ETS DIF Category DSTD-P% F	H	B	A	Item Text
12.	(D) .50 .66	−3.16 C	−1.15 B	−0.29 A	−1.09 B	Noting the potential danger involved in producing nuclear energy as a source of power, many people argue that we should be more systematic in —— the Sun's energy.	(C) .91 .52	−1.65 C	0.52 A	0.48 A	0.88 A	Noting the potential danger involved in producing nuclear energy as a source of power, many people argue that we should be more systematic in —— the Sun's energy.
		0	0	2	0	(A) heating		0	0	−1	0	(A) heating
		11	1	1	4	(B) magnifying		2	1	−1	0	(B) magnifying
		11	0	3	1	(C) dispelling		2	−2	0	−1	(C) dispelling
		−23	−8	−4	−8	*(D) tapping		−5	2	3	2	*(D) utilizing
		0	8	−3	3	(E) discovering		0	0	−1	0	(E) discovering
		1	0	1	−1	(OMITS)		0	1	0	0	(OMITS)
16.	(A) .90 .52	−0.72 A	−1.43 B	−1.31 B	−0.50 A	SCHOOL:FISH::	(B) .92 .32	−0.39 A	−1.37 B	−0.32 A	−1.07 A	HERD:COWS::
		1	0	2	−1	(A) bouquet:flowers		1	5	1	3	(A) bouquet:flowers
		−2	−7	−8	−1	*(B) flock:birds		−1	−5	−1	−4	*(B) flock:birds
		0	2	1	2	(C) crew:ships		0	1	0	1	(C) crew:ships
		0	3	3	0	(D) deluge:water		0	0	0	0	(D) deluge:water
		1	1	3	0	(E) coop:poultry		0	0	1	0	(E) coop:poultry
		0	2	0	0	(OMITS)		0	0	0	0	(OMITS)
16.	(D) .94 .53	−1.43 B	−1.59 B	−0.90 A	−1.39 A	SHORE:LAKE::	(C) .82 .50	−1.22 B	−0.91 A	−1.23 B	−0.54 A	SHORE:LAKE::
		−3	−6	−4	−3	*(A) bank:river		−7	−6	−11	−2	*(A) frame:picture
		1	0	1	2	(B) floor:ocean		2	−1	5	2	(B) floor:ocean
		1	4	0	1	(C) wave:coast		2	0	0	1	(C) wave:coast
		0	1	0	0	(D) height:tower		1	3	1	1	(D) height:tower
		1	2	2	0	(E) current:water		2	4	4	−1	(E) current:water
		0	0	1	0	(OMITS)		1	0	2	0	(OMITS)
17.	(B) .86 .51	−0.32 A	−2.27 C	−2.06 C	−2.37 C	DYE:FABRIC::	(A) .80 .40	0.14 A	−0.55 A	−0.66 A	−0.07 A	DYE:FABRIC::
		0	3	3	3	(A) thinner:paint		−1	4	−1	2	(A) thinner:stain
		0	1	3	4	(B) oil:skin		−3	0	1	0	(B) oil:skin
		−1	−14	−15	−11	*(C) stain:wood		1	−4	−7	0	*(C) paint:wood
		0	2	2	2	(D) fuel:engine		0	3	3	1	(D) fuel:engine
		1	7	7	1	(E) ink:pen		3	−2	3	−3	(E) ink:pen
		0	1	1	1	(OMITS)		0	0	0	0	(OMITS)

MH D-DIF: Mantel-Haenszel Index of Delta Differences (focal minus reference)
ETS DIF Category: A represents negligible DIF, B represents slight to moderate DIF, and C represents moderate to large DIF.
DSTD-P%: Standardization Index of Proportion Correct Differences (focal minus reference)
F: matched female/male comparison H: matched Hispanic/white comparison
B: matched black/white comparison A: matched Asian American/white comparison
*Indicates correct answer.
Item revisions are indicated by boldface.

been postulated to be related to negative DIF for females (Wendler and Carlton 1987). Two "aggression/conflict" items with four versions each are presented in Table 5. Item 13 in Form C was the original pretested item, but as indicated before, this item was no longer classified as C DIF in this investigation. Nevertheless, the three revisions did perform as expected, with the version in Form B showing the least amount of DIF. Unfortunately, the R-Biserial in Form B is .27, slightly below the acceptable level for items (such as this version) of middle difficulty. Item 15, a very difficult sentence completion, remained difficult after each revision (14 to 18 percent correct). The version in Form C changed the context of the sentence from war to economics and did shift the category of negative C DIF to negative B DIF for female examinees, but the change in DSTD value from the original version to that in Form C was only 2 percent, a negligible difference.

Special Interest Terminology

Terminology of special interest or familiarity to a particular group (perhaps due to greater exposure or retention of it by that group) has been hypothesized to affect positively the performance of that group when compared to the performance of a group without this special interest (Schmitt 1985, 1988; Schmitt and Bleistein 1987; Schmitt, Curley, Bleistein, and Dorans 1988; Bleistein, Schmitt, and Curley 1990).

TABLE 9

Mantel-Haenszel Values (MH D-DIF) and ETS DIF Categories for Selected SAT-Verbal Items Reprinted in this Study Identically to the Initial Pretest

Form and Item Number in this Study	*Table Number*	*Focal Group*	*MH D-DIF and ETS DIF Category in this Study*	*MH D-DIF and ETS DIF Category from Prior Pretesting*
Form C, No. 13	5	Females	−1.35 (category B)	−1.80 (category C)
Form A, No. 14	7	Hispanics	−1.19 (category B)	−2.00 (category C)
Form A, No. 16	8	Hispanics	−1.43 (category B)	−1.78 (category C)
Form D, No. 16	8	Hispanics Asian Americans	−1.59 (category B) −1.39 (category A)	−2.04 (category C) −1.91 (category C)
Form D, No. 23*	3	Blacks	−0.83 (category A)	−1.59 (category C)
Form D, No. 25	7	Hispanics	−0.09 (category A)	+1.71 (category C)

*This item was also C DIF for females at initial pretesting and remained C DIF for females when reprinted for this study.

Table 6 presents three item-pairs in which terms deemed of special interest were varied. Items 14 and 24 contain the terms "verve" and "dote," which were considered of possible special interest to female examinees, while item 18 contains the term "plait," which was deemed of special interest to the black group. These judgments about special interests were made *after* analyzing the DIF data from the initial pretesting. In items 14 and 18, only the term considered of special interest was changed in the second version; in item 24, only the two stem terms were changed. In all three items, the revised version (in which the term of special interest was replaced by a hypothetically neutral synonym) was no longer differentially easier for the focal group. No extreme MH C DIF is evident and the DSTD index also indicates a reduction in the expected direction. It is important to note, nevertheless, that the revised versions became notably easier (+25 percent or more) for items 14 and 18.

Cognates

Words that have the same meaning in English as do close approximations of the words in Spanish have been postulated to affect positively the performance of Hispanic examinees when compared to white examinees (Schmitt 1985, 1988; Schmitt, Curley, Bleistein, and Dorans 1988; Schmitt and Dorans 1991).

Table 7 presents two item-pairs in which words considered cognates or noncognates were replaced with synonyms. In item 14, the word "gauge" (in the key) is a noncognate that was replaced in Form B with the word "measure," which is a cognate. Although the reprinted version of this item in Form A is not classified as C DIF as it was when initially pretested, the revision (Form B) does show a reduction in the level of both MH and DSTD DIF for Hispanic examinees (and for females, perhaps because the word "gauge" is used in science and industrial arts). Item 25 was discussed earlier in this paper; it initially showed an elevated C DIF in favor of Hispanics but, when reprinted for this study, no appreciable effect of the change in the stem from "DULCET" to "EUPHONIOUS" appeared.

Homographs

Words spelled and pronounced alike but having multiple meanings have been postulated to be sources of vocabulary confusion that could negatively affect the performance of some focal group examinees when compared to the performance of comparable reference group examinees (Schmitt 1985, 1988; Schmitt and Bleistein 1987; Schmitt, Curley, Bleistein, and Dorans 1988; Bleistein, Schmitt, and Curley 1990; Schmitt and Dorans 1991).

Four item-pairs in which homographs were replaced by comparable terms with single meanings are presented in Table 8. Two of the items (item 16 in Form A and item 16 in Form D) were very easy and, when pretested again, were not classified as extreme C DIF items. Nevertheless, when the homograph was replaced for the other version of these items a small (and insignificant) reduction in the negative MH values was observed. For item 12, the extreme negative DIF observed for female examinees was reduced when the key "tapping" was changed to "utilizing," but the overall difficulty of the item was also reduced considerably. The revised version of the item is still classified as negative C DIF for females but, as discussed earlier, this classification may be an artifact of the MH delta metric. Item 17 behaved almost exactly as expected. The overall item difficulty and discrimination did not change much but, after substituting the word "paint" for "stain" in the key and "stain" for "paint" in the (A) option, the negative DIF observed for all three minority focal groups was reduced considerably.

Conclusions

Several diverse conclusions can be drawn from the data analyzed in this investigation. First, it would appear that revising and re-pretesting SAT-V items to eliminate C DIF is feasible and likely to succeed often enough to make it practical to do so, particularly when prior research on hypothesized DIF factors and/or factors based on observed occurrences of extreme DIF inform the revisions. Changing one or both words in the stems of analogies seemed to be related to the largest and most consistently predictable changes in DIF data, although in many cases such stem revisions also strongly influenced the overall difficulty of the items. Vocabulary-oriented revisions in the keys of sentence completions also proved effective. Item discrimination almost always remained at acceptable levels for the revised versions of both types of questions. Changes only to wrong answer choices (distractors) rarely had strong influence on the DIF data. With the above guidelines in mind—and assuming it is desirable or necessary to do so—it seems appropriate to recommend further such revisions of C DIF items to reduce or eliminate differential difficulty as long as such revised items are re-pretested and reanalyzed for DIF.

Second, the particular terminology used in the stems and keys of analogies and sentence completions seems to be a significant source of elevated levels of DIF on the SAT-V. This hypothesis is supported by the fact that, after the revision of one or two words in most of the items studied, DIF was reduced to acceptable levels. If particular terminology (distinct from underlying analogical reasoning skills or the ability to follow the logic of sentences) is often related to elevated levels of DIF, then evaluation of the construct relevance or irrelevance of individual C DIF items would seem to be appropriate and should be conducted as part of the routine development of tests such as the SAT-V.

Third, to the extent possible, larger sample sizes for focal groups (particularly minority) would seem to be a desirable goal, since the stability of ETS DIF categories is reduced when the sample sizes are small. More than 20 percent of the items studied in this investigation were classified as "C" when first pretested but then as "B" or "A" when reprinted identically for comparable populations. (This percentage is even greater if one considers only the DIF data for the minority groups, for which sample sizes are relatively small.) Such variations are problematic not only because they make it difficult to study the effects of systematic revisions of items but also because, more importantly, they undermine the effort to screen out items with elevated levels of DIF from operational test forms. Without stable classifications, test developers and statisticians cannot be certain which pretest items to review for construct-irrelevant sources of DIF and which pretest items not to review.

Fourth, for classifying the level of DIF (i.e., the "A," "B," and "C" categories), a combination of the Standardization p metric and the Mantel-Haenszel delta metric for very easy and very difficult items seems logical given that the MH (delta-metric) statistic at the extremes of the difficulty continuum has larger standard errors than does the DSTD (p-metric) statistic. Because "the delta metric is unbounded at the extremes..., differences for easy and hard items are played up" (Dorans and Holland 1992, p.27). The fact that DIF data such as those found for item 12 (Form C) in Table 8 and for item 25 (Form A) in Table 4 yield classifications of "C" is unfortunate. These items do not reveal "moderate to large" amounts of DIF; rather, they are merely very easy or very difficult items for which the MH delta metric is not as appropriate an indicator of DIF as is the DSTD p metric.

A final thought relates to the factors (derived from prior DIF research and/or observation of pretested items with extreme DIF) that were used in selecting the items for this investigation. Because the primary purpose of the study was to evaluate whether or not revisions to C DIF items could be made efficaciously, evaluation of the various factors themselves was necessarily ancillary: not many items were studied for most of the individual hypotheses. Yet the authors, in conducting this investigation and attempting to draw conclusions from the data, had to try to determine for themselves the source(s) of the observed C DIF (beyond issues such as sample size, difficulty level, and the metric used). If, as concluded earlier in this section, the particular terminology used in the stems and keys of SAT-V questions is related to elevated levels of DIF, then that DIF is likely also related to reading and other means of vocabulary acquisition, which are part of the construct of reasoning tests such as the SAT-V that measure developed verbal abilities.

It is important that an incorrect "message" not be transmitted to examinees, teachers, and others concerning the application of DIF statistics to the test development process: the deletion of entire categories of items that happen to include some specialized terminology might erroneously suggest that breadth and depth of vocabulary are not important. Students should be encouraged to continue to strive for breadth of coverage in their reading and course work.

Future exploration of the construct relevance of factors related to DIF could address questions such as: Are individual C DIF items that include particular terminology such as that evaluated in this study relevant to the construct measured by tests of developed verbal ability such as the SAT-V? One way to try to answer such a

question empirically would be to determine how particular items or categories of items associated with elevated levels of DIF are related to the predictive validity of such tests for all groups of examinees. Such an exploration could be a significant contribution to future research in this area.

References

Bleistein, C. A., A. P. Schmitt, and W. E. Curley. 1990. "Factors Hypothesized to Affect the Performance of Black Examinees on SAT-Verbal Analogy Items." Paper presented at the annual meeting of the National Council on Measurement in Education, April, Boston, Mass.

Bleistein, C. A., and D. Wright. 1986. "Assessment of Unexpected Differential Difficulty for Asian-American Candidates on the SAT." In *Differential Item Functioning on the Scholastic Aptitude Test* (RM-87-01), ed. A. P. Schmitt and N. Dorans. Princeton, N.J.: Educational Testing Service.

Dorans, N. J. 1989. "Two New Approaches to Assessing Differential Item Functioning: Standardization and the Mantel-Haenszel Method." *Applied Measurement in Education* 2: 217–33.

Dorans, N. J., and P. W. Holland. 1992. *DIF Detection and Description: Mantel-Haenszel and Standardization* (RR-92-10). Princeton, N.J.: Educational Testing Service.

Dorans, N. J., and E. Kulick. 1983. *Assessing Unexpected Differential Item Performance of Female Candidates on SAT and TSWE Forms Administered in December 1977: An Application of the Standardization Approach* (RR-83-9). Princeton, N.J.: Educational Testing Service.

Dorans, N. J., and E. Kulick. 1986. "Demonstrating the Utility of the Standardization Approach to Assessing Unexpected Differential Item Performance on the Scholastic Aptitude Test." *Journal of Educational Measurement* 23: 355–68.

Dorans, N. J., A. P. Schmitt, and C. A. Bleistein. 1992. "The Standardization Approach to Assessing Comprehensive Differential Item Functioning." *Journal of Educational Measurement* 29:309–19.

Dorans, N. J., A. P. Schmitt, and W. E. Curley. 1988. "Differential Speededness: Some Items Have DIF Because of Where They Are, Not What They Are." Paper presented at the annual meeting of the National Council on Measurement in Education, March, New Orleans, La.

Holland, P. W., and D. T. Thayer. 1988. "Differential Item Performance and the Mantel-Haenszel Procedure." In *Test Validity*, ed. H. Wainer and H. I. Braun, pp. 129–45. Hillsdale, N.J.: Erlbaum.

Lawrence, I. M., and W. E. Curley. 1989. *Differential Item Functioning for Males and Females on SAT-Verbal Reading Subscore Items: Follow-up Study* (RR-89-22). Princeton, N.J.: Educational Testing Service.

Lawrence, I. M., W. E. Curley, and F. J. McHale. 1988. *Differential Functioning of SAT-Verbal Reading Subscore Items for Male and Female Examinees* (RR-88-10). Princeton, N.J.: Educational Testing Service.

Mantel, N., and W. M. Haenszel. 1959. "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease." *Journal of the National Cancer Institute* 22: 719–48.

Petersen, N. 1988. "DIF Procedures for Use in Statistical Analysis." Unpublished memorandum issued September 14, 1988.

Rogers, H. J., and E. Kulick. 1987. "An Investigation of Unexpected Differences in Item Performance between Blacks and Whites Taking the SAT." In *Differential Item Functioning on the Scholastic Aptitude Test* (RM-87-01), ed. A. P. Schmitt and N. Dorans. Princeton, N.J.: Educational Testing Service.

Scheuneman, J. D. 1987. "An Experimental Exploratory Study of Causes of Bias in Test Items." *Journal of Educational Measurement* 24: 97–118.

Scheuneman, J. D., and J. A. Briel. 1988. "Differential Effects of Selected Item Factors on the Performance of Hispanic and White Examinees." Paper presented at the annual meeting of the American Educational Research Association, April, New Orleans, La.

Scheuneman, J. D., and K. Gerritz. 1990. "Using Differential Item Functioning Procedures to Explore Sources of Item Difficulty and Group Performance Characteristics." *Journal of Educational Measurement* 27: 109–31.

Schmitt, A. P. 1985. *Assessing Unexpected Differential Item Performance of Hispanic Candidates on SAT Form 3FSA08 and TSWE Form E47* (SR-85-169). Princeton, N.J.: Educational Testing Service.

Schmitt, A. P. 1988. "Language and Cultural Characteristics that Explain Differential Item Functioning for Hispanic Examinees on the Scholastic Aptitude Test." *Journal of Educational Measurement* 25: 1–13.

Schmitt, A. P., and C. A. Bleistein. 1987. *Factors Affecting Differential Item Functioning for Black Examinees on Scholastic Aptitude Test Analogy Items* (RR-87-23). Princeton, N.J.: Educational Testing Service.

Schmitt, A. P., W. E. Curley, C. A. Bleistein, and N. J. Dorans. 1988. "Experimental Evaluation of Language and Interest Factors Related to Differential Item Functioning for Hispanic Examinees on the SAT-Verbal." Paper presented at the annual meeting of the National Council on Measurement in Education, March, New Orleans, La.

Schmitt, A. P., and N. J. Dorans. 1990. "Differential Item Functioning for Minority Examinees on the SAT." *Journal of Educational Measurement* 27: 67–81.

Schmitt, A. P., and N. J. Dorans. 1991. "Factors Related to Differential Item Functioning for Hispanic Examinees on the Scholastic Aptitude Test." In *Assessment and Access:*

Hispanics in Higher Education, ed. G. D. Keller, J. R. Deneen, and R. J. Magallan, pp. 105–32. New York: SUNY Press.

Wendler, C. L. W., and S. T. Carlton. 1987. "An Examination of SAT-Verbal Items for Differential Performance by Women and Men: An Exploratory Study." Paper presented at the annual meeting of the American Educational Research Association, April, Washington, D.C.

Wright, D. 1987. "An Empirical Comparison of the Mantel-Haenszel and Standardization Methods of Detecting Differential Item Performance." In *Differential Item Functioning on the Scholastic Aptitude Test* (RM-87-01), ed. A. P. Schmitt and N. Dorans. Princeton, N.J.: Educational Testing Service.

Zieky, M. 1991. "Using DIF Statistics in TD: Practical Issues." Paper presented at the annual meeting of the National Council on Measurement in Education, April, Chicago, Ill.

Appendix

Summary of Hypotheses about DIF Relevant to the SAT-Verbal Items Selected for this Study

Description of Hypothesis (and References, if any)*	*Total Number of Items Studied*	*Table of Items and Data*
Technical/specialized science terminology may negatively affect the performance of females (Lawrence, Curley, and McHale 1988; Lawrence and Curley 1989; Scheuneman and Gerritz 1990)	4	Table 2
Technical/specialized industrial arts terminology may negatively affect the performance of females (no references from research—based on empirical observation of SAT-V pretest results)	2	Table 3
Technical/specialized military terminology may negatively affect the performance of females (no references from research—based on empirical observation of SAT-V pretest results)	6	Table 4
Contexts portraying aggression or conflict may negatively affect the performance of females (Wendler and Carlton 1987)	2	Table 5
Terminology of special interest or familiarity to a group may positively affect the performance of that group (Schmitt 1985, 1988; Schmitt and Bleistein 1987; Schmitt, Curley, Bleistein, and Dorans 1988; Bleistein, Schmitt, and Curley 1990)	3	Table 6
Cognates with Spanish may positively affect the performance of Hispanic examinees (Schmitt 1985, 1988; Schmitt, Curley, Bleistein, and Dorans 1988; Schmitt and Dorans 1991)	2	Table 7
Homographs may negatively affect the performance of Hispanic, black, and Asian American examinees (Schmitt 1985, 1988; Schmitt and Bleistein 1987; Schmitt, Curley, Bleistein, and Dorans 1988; Bleistein, Schmitt, and Curley 1990; Schmitt and Dorans 1991)	4	Table 8

*Not all (or even most) SAT-Verbal items in these seven general categories consistently show elevated levels of DIF, but certain patterns have been detected.